INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2022-2024

FGAI4H-P-009-A02

ITU-T Focus Group on AI for Health

Original: English

WG(s):	Plen	Helsinki, 20-22 September 2022

DOCUMENT

Source:	TG-DiagnosticCT Topic Driver
Title:	Att.2 – CfTGP update (TG-DiagnosticCT) [same as Meeting H]
Purpose:	Engagement

Contact: Kuan Chen Email: ckuan@infervision.com
InferVision, China

Abstract: Calling on members of the medical and artificial intelligence communities with a vested interest in AI for volumetric chest computed tomography (CT)! Become engaged in the group dedicated to establishing a standardized benchmarking platform for AI for volumetric chest CT within the International Telecommunication Union (ITU)/World Health Organization (WHO) Focus Group on "Artificial Intelligence for Health" (FG-AI4H).

This version of the CfTGP is the same as seen in Meeting H (FGAI4H-H-009-A02), reproduced for easier reference as a Meeting N document.

Call for Topic Group Participation: AI for Volumetric Chest CT

The International Telecommunication Union (ITU)/World Health Organization (WHO) Focus Group on "Artificial Intelligence for Health" (FG-AI4H; https://www.itu.int/go/fgai4h) seeks engagement from members of the medical and artificial intelligence (AI) communities (including clinicians, technologists, entrepreneurs, potential benchmarking data providers, machine learning experts, software developers, researchers, regulators, policy-makers, companies/institutions, and field experts) with a vested interest in shaping the benchmarking process of AI for volumetric chest CT.

About FG-AI4H

Over the past decade, considerable resources have been allocated to exploring the use of AI for health, which has revealed an immense potential. Yet, due to the complexity of AI models, it is difficult to understand their strengths, weaknesses, and limitations. If the technology is poorly designed or the underlying training data are biased or incomplete, errors or problematic results can occur. AI technology can only be used with complete confidence if it has been quality controlled through a rigorous evaluation in a standardized way. Towards developing this standard assessment framework of AI for health, the ITU has established FG-AI4H in partnership with the WHO.

Thus far, FG-AI4H has established thirteen topic groups. These are concerned with: AI and cardiovascular disease risk prediction, child growth monitoring, dermatology, falls among the elderly, histopathology, neuro-cognitive diseases, ophthalmology (retinal imaging diagnostics), psychiatry, radiotherapy, snakebite and snake identification, symptom checkers, tuberculosis, and volumetric chest computed tomography.

Each topic group agrees upon representative benchmarking tasks in a pragmatic, best-practice approach, which can later be scaled and expanded to similar tasks. Every benchmarking task should address a health problem of relevance (e.g. impacting a large and diverse part of the global population or challenging to treat) and for which AI technology would provide a tangible improvement relative to the current practice (e.g. better care, results, and/or cost/time effectiveness).

For a rigorous and sound evaluation, undisclosed test data sets must be available (or have to be collected) for each task. All data must be of high quality and compliant with ethical and legal standards. In addition, the data must originate from a variety of sources so that it can be determined whether an AI algorithm can generalize across different conditions, locations, or settings (e.g. across different people, hospitals, and/or measurement devices). The format/properties of the data serving as input to the AI and of the output expected from the AI, as well as the benchmarking metrics are agreed upon and specified by the topic group.

Finally, the AI-to-be-evaluated will be benchmarked with the undisclosed test data on FG-AI4H computing infrastructure. Here, the AI will process single samples of the undisclosed test data set and predict output variables, which will be compared with the "ground truth." The results of the benchmarking will be provided to the AI developers and will appear on a (potentially anonymized) leaderboard.

Topic group: AI for Volumetric Chest CT

1) A topic group is a community of stakeholders from the medical and AI communities with a shared interest in a topic. The objectives of the topic groups are manifold: to provide a forum for open communication among various stakeholders,
2) to agree upon the benchmarking tasks of this topic and scoring metrics,
3) to facilitate the collection of high quality labelled test data from different sources,
4) to clarify the input and output format of the test data,
5) to define and set-up the technical benchmarking infrastructure, and
6) to coordinate the benchmarking process in collaboration with the Focus Group management and working groups.

The primary output of a topic group is one document that describes all aspects of how to perform the benchmarking for this topic. (The document will be developed in a cooperative way by suggesting changes as input documents for the next FG-AI4H meeting that will then be discussed and integrated into an official output document of this meeting. The process will continue over several meetings until the topic description document is ready for performing the first benchmarking.)

This topic group is dedicated to AI for volumetric chest CT.

1 Introduction

1.1 Topic and Relevance

According to a World Health Organization (WHO) cancer report, lung cancer remains the leading cause of morbidity and mortality with an estimated 2.09 million new cases and 1.76 million death worldwide in 2018. The overall five-year survival rate is merely 10-15%, early detection is crucial because stage I lung cancer is reported to have a 10-year survival rate of ~90% if surgical resections were conducted. Early lung cancer characterization (including differentiating benign from malignant nodules) is also critical because this would dramatically affect treatment selection and patient outcome. Annual screening by low-dose volumetric chest computed tomography has thus been recommended by National Lung Screening Trial (NLST) for early lung cancer detection to

reduce lung cancer mortality. However, large numbers of thin slices (300-400) were generated in one lung CT scan and the interpretation is tedious and time-consuming and small nodules are often overlooked while misdiagnoses are not uncommon. In China, lung cancer is the leading cause of mortality with a recorded number of 610,200 lung cancer deaths in 2015 and it is projected that lung cancer mortality in China may increase by approximately 40% between 2015 and 2030.

The two biggest issues in Chinese medical system are shortage and centralization of medical resources. There is a serious shortage of qualified radiologists with an annual growth rate of ~4% while the medical imaging demand increase rate was nearly 30%. Patients tend to seek medical care in large tertiary hospitals in top tier cities while community hospitals, village health centers and clinics are largely underused. As a result, the workloads of radiologists were immensely intensified especially for big hospitals since they must examine hundreds of images every day while rural city radiologists remain less skilled which make early lung cancer detection and characterization more difficult.

1.2 Impact of AI-assisted Diagnostic System

AI which mimics the function of human brain is a promising method that has been demonstrated to possess diagnostic performance equal or even surpass clinical experts. AI could increase diagnostic accuracy and lower physician burnouts.

2 Intended Benchmark Tasking

2.1 Clinical Requirement

Lung nodule detection refers to the identification of “a round opacity at least moderately well marginated measuring less than 3 cm” according to the Fleischner society guideline while nodule characterization refers to confirmation of nodule property such as size, density, shape, margin, internal structure, adjacent structure, location, relationship to vascular, lung hilum or mediastinum structures, benign or malignant which greatly affect treatment selection and patient outcome. Nodule detection is as important as nodule characterization.

2.2 AI Algorithm Testing

2.2.1 Fundamental Principles

It is necessary that the model is robust enough that minimal/no model tuning is allowed for testing. When evaluating an AI algorithm, all parameters of the model should be fixed prior to the test and there should be no change to the algorithm.

2.2.2 Available Public and Undisclosed Test Data

1) In order to assess algorithm robustness, sufficient and diversified data from multiple heterogeneous sources (e.g., imaging device models, imaging protocols, patient demographics, clinical conditions, etc.) should be used for testing to verify the generalization capacity of AI. Public and real-world undisclosed data (desensitized) should be enrolled.
2) Public database may include Lung Image Database Consortium Image Collection from Cancer Imaging Archive (LIDC/LDRI) and NLST.
3) To avoid prior finetuning and overfitting on the public data by algorithms, mostly real-world undisclosed data shall be used for the test.
4) Undisclosed test data should include geographic differences (terrain such as plain and plateau), gender differences, multiple countries (from >3 continents) with different socioeconomics and

the resulting lifestyles (>=5 countries), multiple hospitals (>=20) including big hospitals (>800 beds) and small hospitals and different races.

5) Undisclosed test data should come from various CT vendors (GE, Siemens, Philips, Canon, United Imaging, etc.) reconstructed with various thin slices (0.625mm, 1mm, 1.25mm, 2mm, 3mm, etc.) and thick slices (5mm, 7mm, etc.). Various reconstruction kernels such as lung kernel, bone kernel, standard kernel should also be included.

6) Undisclosed data should include low dose CT scans (with effective does of 1.5~2.5 mSv) and normal dose CT scans (2.5 mSv~5 mSv).

7) All testing data would preferably have pathological finding confirmation of nodule existence and property.

8) Testing data should include pulmonary nodules of different types such as varying sizes, varying densities (solid, part-solid, pure ground-glass), varying shapes (round, spiculated, lobulated, etc.) to test the detection capability of AI on different nodule types.

9) Testing data should include certain number of negative cases (without nodules or with lesions exhibiting similar manifestations of nodules) to test the false positive rate of AI.

10) It would be good for undisclosed test data pool size be as large as possible (preferably >5, 000 cases)

11) Case-control subsets of testing data (subtle cases) should also be included for comprehensive evaluation of AI performance in various perspectives. Several subsets were recommended: e.g. easily missed but important lesions like GGN, nodules in close proximity with vessels, nodules in close proximity to cavity etc.

12) An arbitration commission compromising of chest CT experts will examine labeled undisclosed test patient data to confirm data sufficiency, heterogeneity, labelling quality and conformity to ethical and legal standards.

2.2.3 Gold Standard Labelling of Test Data

1) For clinical evaluation of AI algorithm for lung nodule detection, gold standard labelling is obtained from a panel of experts. The panel should involve at least three experts with 10 years of practice in thoracic radiology or above. One of the experts serves as an arbitrator to make final decisions on disputed annotations from the other experts. All labelling experts would receive prior specialized training regarding how to annotate nodules, delineation of the nodule boundaries, long-axis diameter at maximum cross section measurement, etc.

2) For clinical evaluation of AI for malignant nodule diagnosis, gold standard comes from pathological or CT follow-ups. As for pathological results, both pathological type, stage and pathology details (whether obtained by liquid biopsy or fine needle aspiration biopsy, whether confirmed by histopathology or cytology) should be recorded.

3) The gold standard annotation of testing data should be strictly monitored and managed in order to preserve the confidentiality of gold standard testing data results.

2.2.4 Standalone Testing

In the standalone test which evaluates AI's performance independently, influencing factors such as nodule size, nodule type, scanning and image reconstruction protocols, imaging devices, population characteristics were tested by dividing testing data into corresponding category groups. The participants submit the algorithms through an online portal, after which the algorithms will be run against the undisclosed dataset. Inputs and outputs of the test will be defined as follow.

2.2.5 AI Benchmarking Input and Output

Chest CT volumetric CT images obtained in first scans or follow-ups were eligible for AI benchmarking. Original DICOM data together with labelling information were required. Image resolution should be no less than 512*512 pixels. The AI output should include information about lesions position, long axis diameter, classification, density and malignancy probability.

2.2.6 Primary Benchmarking Metrics

Primary testing includes nodule detection with overall dataset (including public data and undisclosed data). Detected nodules by standalone AI system, were compared with pre-labelled nodules to determine the true positive nodules and false positive nodules, benchmarking metrics including nodule-based sensitivity, false positive rate, free-response ROC, location ROC, AFORC,F-score, time spent reading the cases were calculated. Detection capability of AI on different nodule types such as varying sizes, varying densities (solid, part-solid, pure ground-glass), varying shapes (round, spiculated, lobulated, etc.) were calculated.

For malignancy classification task, sensitivity, specificity, ROC, time spent reading the cases were calculated.

2.2.7 Secondary Benchmarking Metrics

In addition to the primary benchmarking metrics, other performance metrics relevant to the task at hand should be included as secondary benchmark metrics.

1) Nodule characterization: size measurement, density, classification and malignancy

   In nodule size measurement, error rate which is the absolute value of error between AI measured nodule long axis diameter and radiologist's labelling nodule long axis diameter divided by radiologist's labelling diameter should be calculated. In nodule density characterization (which classify nodules into solid, part-solid, ground glass opacity and calcified nodules), accuracy (defined as true positive nodules divided by gold standard nodules) was used to depict nodule characterization capability. For malignancy characterization, nodules were classified as benign and malignant, sensitivity, specificity, accuracy and AUC were used to evaluate testing performance.

2) Subtle case performances: e.g. easily missed but important lesions like GGN, nodules in close proximity to vessels, nodules in close proximity to cavity etc.

   Subtle cases are subsets of the test data which held great clinical significance and shall be tested independently. Subtle cases should be selected from the original pool of testing data, preferably including more than 100 cases or more for each subset. Performances metrics can be adopted from primary benchmarking.

2.2.8 Testing Process Control

Testing personnel and procedure details should be included. Testing data should be kept untouched to AI companies under all circumstances. Physician annotations should be conducted in WHO/ITU certified centers kept confidential at all time.

More details about the activities of the topic group can be found in the document D-024, which can be accessed with a free ITU account (cf. "Get involved").

The topic group on AI for volumetric chest CT is led by Kuan Chen (InferVision, China) who encourages interested parties to join as topic group members. The topic group would benefit from further expertise of the medical and AI communities and from additional data.

1 Get involved

To join this topic group, please send an e-mail to the focus group secretariat (tsbfgai4h@itu.int) and to Pei Zhang(zpei@infervision.com) (as representative of the topic driver Kuan Chen). Please use a descriptive e-mail subject (e.g. "Participation topic group AI for volumetric chest CT"), briefly introduce yourself and your organization, concisely describe your relevant experience and expertise, and explain your interest in the topic group.

Participation in FG-AI4H is free of charge and open to all. To attend the workshops and meetings, please visit the Focus Group website (https://www.itu.int/go/fgai4h), where you can also find the whitepaper, get access to the documentation, and sign up to the mailing list.

_____